

Formal Analysis of Damasio's Theory on Core Consciousness

Tibor Bosse (tbosse@cs.vu.nl)

Vrije Universiteit Amsterdam, Department of Artificial Intelligence
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

Catholijn M. Jonker (C.Jonker@nici.ru.nl)

Radboud Universiteit Nijmegen, Nijmegen Institute for Cognition and Information
Montessorilaan 3, 6525 HR Nijmegen, The Netherlands

Jan Treur (treur@cs.vu.nl)

Vrije Universiteit Amsterdam, Department of Artificial Intelligence
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

Abstract

This article presents a formal analysis of Damasio's theory on core consciousness. Three important concepts in this theory are "emotion", "feeling", and "feeling a feeling" (or core consciousness). In particular, a simulation model is described of the neural dynamics leading via emotion and feeling to core consciousness, and dynamic properties are formally specified that hold for these dynamics. These properties have been automatically checked for the simulation model. Moreover, a formal analysis is made and verified of relevant notions of representation.

Introduction

The neurologist Damasio (2000) describes his theory of consciousness in an informal manner, and supports it by evidence from neurological practice. More experimental work supporting his theory is reported in (Damasio *et al.*, 2000; Parvizi & Damasio, 2001). Damasio's theory is described on the one hand in terms of the occurrence of certain neural states (or neural patterns), and temporal or causal relationships between them. Formalisation of these relationships requires a modelling format that is able to express direct temporal or causal dependencies. On the other hand, Damasio gives interpretations of most of these neural states as representations, for example as 'sensory representation', or 'second-order representation'. This requires an analysis of what it means that a neural state is a representation for something. This paper focuses on Damasio's notions of 'emotion', 'feeling', and 'core consciousness' or 'feeling a feeling'. Damasio (2000) describes an *emotion as neural object* (or *internal emotional state*) as an (unconscious) neural reaction to a certain stimulus, realized by a complex ensemble of neural activations in the brain. As the neural activations involved often are preparations for (body) actions, as a consequence of an internal emotional state, the body will be modified into an *externally observable emotional state*. Next, a *feeling* is described as the (still unconscious) sensing of this body state. Finally, *core consciousness* or *feeling a feeling* is what emerges when the organism detects that its representation of its own body state (the *proto-self*) has been changed by the

occurrence of the stimulus: it becomes (consciously) aware of the feeling.

This paper aims at formalisations and simulation models for these three notions. In addition, the notion of representation used by Damasio is formally analysed against different approaches to representational content from the literature on the Philosophy of Mind. It is shown that the classical causal/correlational approach to representational content (e.g., Kim, 1996, pp. 191-193) is inappropriate to describe the notion of representation for core consciousness used by Damasio, as this notion essentially involves more complex temporal relationships describing histories of the organism's interaction with the world. An alternative approach is shown to be better suited: representational content as relational specification over time and space (cf. Kim, 1996, pp. 200-202). Criteria for this approach are formalised, and it is shown that the formalisation of Damasio's notions indeed fit these criteria.

In this paper, first the modelling approach used is briefly introduced. Next, for a simple example, formal models are presented for the processes leading to emotion, feeling, and feeling a feeling (or conscious feeling), and simulation results of these models are shown. After that, it is analysed to what extent the representational content of Damasio's notions can be described by two approaches from Philosophy of Mind. Formalisations of some of the dynamic properties of the processes leading to emotion, feeling and feeling a feeling are presented. Finally, the issue of verification is addressed: it is shown that the developed notions for representational content indeed hold for the simulation model. The verification is performed both by automated checks and by mathematical proof.

Modelling Approach

To model the making of emotion, feeling and core consciousness, dynamics play an important role. Dynamics will be described in the next section as evolution of *states* over time. The notion of state as used here is characterised on the basis of an ontology defining a set of state properties that do or do not hold at a certain point in time. The modelling perspective taken is not a symbolic perspective, but essentially addresses the neural processes and their

dynamics as neurological processes. This implies that states are just neurological states. To successfully model such complex processes, forms of abstraction are required; e.g.:

- neural states or activation patterns are modelled as single state properties
- large-dimensional vectors of such (distributed) state properties are composed to one single composite state property, when appropriate; e.g., (p1, p2, ...) to p and (S1, S2, ...) to S in the next section

To describe the dynamics of the processes mentioned above, explicit reference is made to time. Dynamic properties can be formulated that relate a state at one point in time to a state at another point in time. A simple example is the following dynamic property specification for belief creation based on observation:

‘at any point in time t1, if the agent observes rain at t1, then there exists a point in time t2 after t1 such that at t2 the agent has internal state property s’

Here, for example, s can be viewed as a sensory representation of the rain. To express dynamic properties in a precise manner a language is used in which explicit references can be made to time points and traces: the Temporal Trace Language TTL (cf. Jonker & Treur, 2002). Here a *trace or trajectory* over an ontology Ont is a time-indexed sequence of states over Ont. The sorted predicate logic temporal trace language TTL is built on atoms referring to, e.g., traces, time and state properties. For example, ‘in the internal state of agent A in trace γ at time t property s holds’ is formalised by $\text{state}(\gamma, t, \text{internal}(A)) \models s$. Here \models is a predicate symbol in the language, usually used in infix notation, which is comparable to the Holds-predicate in situation calculus. Dynamic properties are expressed by temporal statements built using the usual logical connectives and quantification (for example, over traces, time and state properties).

To be able to perform some (pseudo)-experiments, a simpler temporal language has been used to specify simulation models in a declarative manner. This language (the *leads to* language) enables modelling direct temporal dependencies between two state properties in successive states. This executable format is defined as follows. Let α and β be state properties of the form ‘conjunction of atoms or negations of atoms’, and e, f, g, h, non-negative real numbers. In *leads to* the notation $\alpha \rightarrow_{e, f, g, h} \beta$, means:

If state property α hold for a time interval with duration g, then after some delay (between e and f) state property β will hold for a time interval of length h.

For a precise definition of the *leads to* format in terms of the language TTL, see (Jonker, Treur & Wijngaards, 2003). A specification of dynamic properties in *leads to* format has as advantages that it is executable and that it can often easily be depicted graphically.

Emotion

First Damasio’s notion of *emotion* is addressed. He describes an *internal emotional state* is a collection of neural dispositions in the brain, which are activated as a reaction on a certain stimulus. Once such an internal

emotional state occurs, it entails modification of both the body state and the state of other brain regions. By these events, an *external emotional state* is created, which is accessible for external observation.

Assume that the music you hear is so special that it leads to an emotional state in which you show some body responses on it (e.g., shivers on your back). This process is described by executable local dynamic properties taking into account an internal state property for activated sensory representation of hearing the music (modelled by sr(music)), and a vector for the activation of preparatory states (p1, p2, ...) for the body responses (S1, S2, ...); see Figure 1.

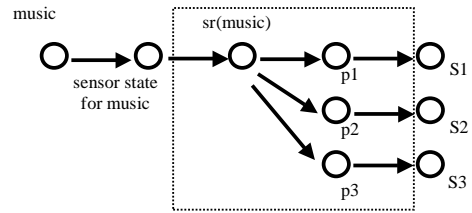


Figure 1: Processes leading to an emotional state.

These vectors are the possible internal emotional states. Note that the state properties are abstract in the sense that a state property refers to a specific neural activation pattern. In the model the conjunction p1 & p2 & .. of these preparatory state properties is denoted by p; this p can be considered a composite state property. Moreover, the conjunction of the vector of all body state properties responding to the music S1, S2, ... (i.e., the respective body state properties for which p1, p2, ... are preparing) is denoted by (composite) state property S.

The model abstracted in this manner is depicted in Figure 2, upper part. In formal textual format these local properties (LP’s) are as follows:

- LP0 music \leftrightarrow sensor_state(music)
- LP1 sensor_state(music) \leftrightarrow sr(music)
- LP2 sr(music) \leftrightarrow p
- LP3 p \leftrightarrow S

In the remainder of this paper this abstract type of modelling will be used. Notice, however, that each of the abstract state properties used are realised in the organism in a distributed manner as a large-dimensional vector of more local (neural) state properties. Also the sensory representation sr(music) may be considered such a composite state property with different aspects of the music represented in different forms at different places. Notice, moreover, that the names of the state properties have been chosen to support readability for humans. But in principle these names should be considered neutral indications of neural states, such as n1, n2, and so on.

Feeling

Next, Damasio’s notion of *feeling* is considered. According to Damasio, a feeling emerges when the collection of neural patterns contributing to the emotion lead to mental images. In other words, the organism senses the consequences of the

internal emotional state. Damasio distinguishes two mechanisms by which a feeling can be achieved:

(1) Via the *body loop*, the internal emotional state leads to a changed state of the body, which subsequently, after sensing, is represented in somatosensory structures of the central nervous system.

(2) Via the *as if body loop*, the state of the body is not changed. Instead, on the basis of the internal emotional state, a changed representation of the body is created directly in sensory body maps. Consequently, the organism experiences the same feeling as via the body loop: it is ‘as if’ the body had really been changed but it was not.

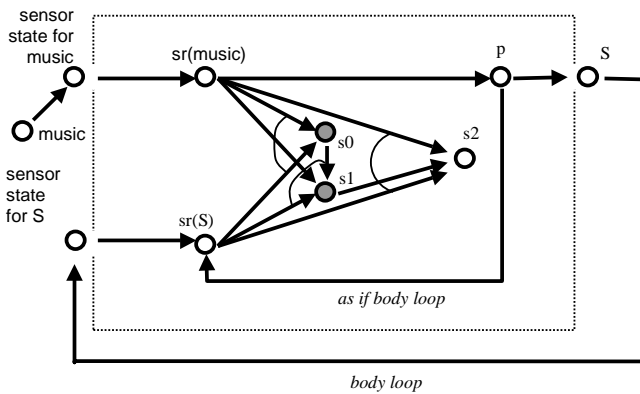


Figure 2: Overview of the simulation model.

The model described in the previous section can be extended to include a number of internal state properties for sensory representations of body state properties that are changed due to responses on the music; together these sensory representations constitute the feeling induced by the music. In Figure 2 the conjunction of these sensory representations is depicted: sr(S) (a sensory representation of the changed body state; this may be materialised in a distributed manner as a kind of vector). This describes the ‘body loop’ for the responses on the music; here S and sensor_state(S) are effects and sensors in the body, respectively. In formal format, two additional local dynamic properties are needed (see also Figure 2):

LP4 $S \leftrightarrow \text{sensor_state}(S)$

LP5 $\text{sensor_state}(S) \leftrightarrow \text{sr}(S)$

Notice that an internal state property sr(shivering) for shivering only, does not directly relate to the music. It is caused by the external stimulus shivering, which in this particular case is originally caused by the music. This body state property shivering could be present for a lot of other reasons as well, e.g., a cold shower. However, taking into account that not only shivering but a larger number of sensory state properties constitute the overall composite state property sr(S), the feeling will be more unique for the music. For the case of an ‘as if body loop’ dynamic properties LP3, LP4 and LP5 can be replaced by the following local dynamic property directly connecting p and sr(S).

LP6 $p \leftrightarrow \text{sr}(S)$

Also a combination of models can be made, in which some effects of hearing the music is caused by a body loop and some are caused by an ‘as if body loop’.

Feeling a Feeling

Finally, Damasio’s notion of *knowing* or *being conscious of* or *feeling a feeling* is addressed. This notion is based on the organism detecting that its representation of its own (body) state (the *proto-self*) has been changed by the occurrence of a certain object (the music in our example). According to Damasio, the proto-self is “a coherent collection of neural patterns which map, moment by moment, the state of the physical structure of the organism”. (Damasio, 2000, p. 177). The conscious feeling occurs when the organism detects the transitions between the following moments:

1. The proto-self exists at the inaugural instant.
2. An object comes into sensory representation.
3. The proto-self has become modified by the object.

For our case we restrict ourselves to placing the relevant events in a temporal context. In a detailed account, in the trace considered subsequently the following events take place: no sensory representations for music and S occur, the music is sensed, the sensory representation sr(music) is generated, the preparation representation p for S is generated, S occurs, S is sensed, the sensory representation sr(S) is generated. According to Damasio (2000, pp. 177-183), two transitions are relevant (see Damasio’s Figure 6.1), and have to be taken into account in a model:

§ from the sensory representation of the initial no S body state and not hearing the music to hearing music and a sensory representation of the music, and no S sensory representation

§ from a sensory representation of the music and no sensory representation of S to a sensory representation of S and a sensory representation of the music

These two transitions are to be detected and represented by the organism. To model this process three internal state properties are introduced: s0 for encoding the initial situation, and s1 and s2 subsequently for encoding the situations after the two relevant changes. By making these state properties persistent they play the role of indicating that in the past a certain situation has occurred. Local dynamic properties that relate these additional internal state properties to the others can be expressed as follows (see also Figure 2); here state properties s0 and s1 are persistent:

LP7 $\text{not sr}(\text{music}) \ \& \ \text{not sr}(S) \leftrightarrow s0$

LP8 $\text{sr}(\text{music}) \ \& \ \text{not sr}(S) \ \& \ s0 \leftrightarrow s1$

LP9 $\text{sr}(\text{music}) \ \& \ \text{sr}(S) \ \& \ s1 \leftrightarrow s2$

Simulation

A special software environment has been created to enable the simulation of executable models (Bosse *et al.*, 2005). Based on an input consisting of dynamic properties in *leads to* format (and their timing parameters e, f, g, h), this software environment generates simulation traces. Using this software environment, the model described in the

previous sections has been used to generate a number of simulation traces. An example of such a simulation trace can be seen in Figure 3. Here, time is on the horizontal axis, the state properties are on the vertical axis. A dark box on top of the line indicates that the property is true during that time period, and a lighter box below the line indicates that the property is false. This trace is based on all executable local properties (i.e., LP0 to LP9), except LP6. In all properties, the values (0,0,1,1) have been chosen for the timing parameters e , f , g , and h . Figure 3 shows how the presence of the music first leads to an emotion (p or S), then to a feeling ($sr(S)$), and finally to the birth of core consciousness ($s2$), involving a body loop.

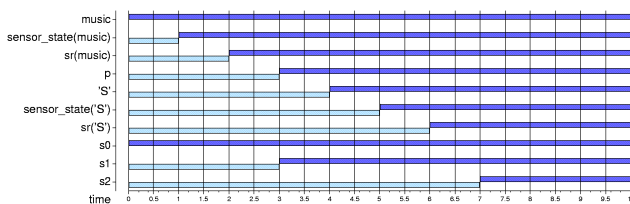


Figure 3: Simulation trace involving a body loop.

Representational Content

In Damasio's description various types of representation are used, for example, sensory representations and second-order representations. In the literature on Philosophy of Mind a number of approaches to representational content are discussed.

In (Kim, 1996, pp. 191-192) the *causal/correlational approach* to representational content is explained as follows. Suppose that, some causal chain is connecting an internal state property s and external state property 'horse nearby'. Due to this causal chain, under normal conditions internal state property s of an organism covaries regularly with the presence of a horse: this state property s occurs precisely when a horse is present nearby. Then the occurrence of s has the presence of the horse as its representational content. Especially for perceptual state properties this may work well.

In (Kim, 1996, pp. 200-202) the concept of *relational specification* of a state property is put forward as an approach to representational content. It is based on a specification of how an internal state property can be related to properties of states distant in space and time. This approach is more liberal than the causal/correlational approach, since it is not restricted to one external state, but allows reference to a whole sequence of states in history.

In the following sections it is explored whether these approaches can be used to specify the representational content of the relevant mental states that occur in our model (i.e., the states that represent emotion, feeling, and feeling a feeling).

Content of Emotion

Consider the causal chain $music - sensor_state(music) - sr(music) - p - S$ (see Figure 1). Thus, looking backward in

time, the external emotional state property S can be considered to (externally) represent the emotional content of the music. On the other hand, the internal emotional state property involved is p . Given the causal chain above the (backward) representational content for both p and S is the presence of this very special music, which could be considered acceptable. However, following the same causal chain, also the state property $sr(music)$ has the same representational content. What is different between p and $sr(music)$? Why are the emotional responses to the same music different between different individuals? This would not be explainable if in all cases the same representational content is assigned. It might be assumed that state properties such as $sr(music)$ may show changes between different individuals. However, the differences are probably much larger between the ways in which for two different individuals $sr(music)$ is connected to a composite state property p . This subjective aspect is not taken into account in the causal/correlational approach. The content of such an emotional response apparently is more personal than a reference to an objective external factor, so to define this representational content both the external music and the internal personal make up has to be taken into account.

For the relational specification approach the representational content of p can be specified in a manner similar to the causal/correlational approach by 'p occurs if the very special music just occurred', and conversely. However, other, more suitable possibilities are available as well, such as, 'p occurs if the very special music just occurred, and by this organism such music was perceived as $sr(music)$ and for this organism $sr(music)$ leads to p ', and conversely. This relational specification involves both the external music and the internal make up of the organism, and hence provides a subjective element in the representational content, in addition to the external reference. This provides an explanation of differences in emotional content of music between individuals.

Content of Feeling

The representational content of $sr(S)$ according to the causal/correlational approach can consider the causal chain $music - sensor_state(music) - sr(music) - p - S - sensor_state(S) - sr(S)$. Using this chain, $sr(S)$ can be related to both the presence of S , and further back to the presence of the very special music. This steps outside the context of having a reference to one state, which limits the causal/correlational approach. A more suitable approach is the relational specification approach, which allows such temporal relationships to different states in the past; there is the following temporal relation between the occurrence of $sr(S)$, the presence of the S , and the presence of music: ' $sr(S)$ occurs if S just occurred, preceded by the presence of the music', and conversely.

Content of Feeling a Feeling

The representational content of $s0$ according to the causal/correlational approach can be taken as the absence of both S and music in the past, via the causal chain: no S and no music - sensor state no S and sensor state no music - no

sr(music) and no sr(S) - s0. This can be expressed relationally by referring to one state in the past: ‘if no S and no music occur, then later s0 will occur,’ and conversely. Formally:

$$\begin{aligned} \forall t1 \ [\text{state}(\gamma, t1, EW) \models \neg S \wedge \neg \text{music} \Rightarrow \\ \exists t2 \geq t1 \ \text{state}(\gamma, t2, \text{internal}) \models s0] \\ \forall t2 \ [\text{state}(\gamma, t2, \text{internal}) \models s0 \Rightarrow \\ \exists t1 \leq t2 \ \text{state}(\gamma, t1, EW) \models \neg S \wedge \neg \text{music}] \end{aligned}$$

For s1 and s2 the causal/correlational approach does not work very well because these state properties essentially encode (short) histories of states. For example, the representational content of s1 according to causal/correlational approach can be tried as follows: presence of the music and no S in the past under the condition that at some point in time before that point in time no music occurred. However, this cannot be expressed adequately according to the causal/ correlational approach since it is not one state in the past to which reference is made, but a history given by some temporal sequence. The problem is that no adequate solution is possible, since the internal state properties should in fact be related to sequences of different inputs over time in the past. This is something the causal/correlational approach cannot handle, as reference has to be made to another state at one time point, and it is not possible to refer to histories, i.e., sequences of states over time, in the past. A better option is provided by representational content of s1 as relational specification: ‘if no S and no music occur, and later music occurs and still no S occurs, then still later s1 will occur,’ and conversely. Formally:

$$\begin{aligned} \forall t1, t2 \ [\ t1 \leq t2 \ \& \\ \text{state}(\gamma, t1, EW) \models \neg S \wedge \neg \text{music} \ \& \\ \text{state}(\gamma, t2, EW) \models \neg S \wedge \text{music} \ \Rightarrow \\ \exists t3 \geq t2 \ \text{state}(\gamma, t3, \text{internal}) \models s1] \\ \forall t3 \ [\ \text{state}(\gamma, t3, \text{internal}) \models s1 \ \Rightarrow \exists t1, t2 \ \ t1 \leq t2 \leq t3 \ \& \\ \text{state}(\gamma, t1, EW) \models \neg S \wedge \neg \text{music} \ \& \\ \text{state}(\gamma, t2, EW) \models \neg S \wedge \text{music} \] \end{aligned}$$

Similarly, the representational content of s2 as relational specification can be specified as follows: ‘if no S and no music occur, and later music occurs and still no S occurs, and later music occurs and S occurs, then still later s2 will occur,’ and conversely. Formally:

$$\begin{aligned} \forall t1, t2, t3 \ [\ t1 \leq t2 \leq t3 \ \& \\ \text{state}(\gamma, t1, EW) \models \neg S \wedge \neg \text{music} \ \& \\ \text{state}(\gamma, t2, EW) \models \neg S \wedge \text{music} \ \& \\ \text{state}(\gamma, t3, EW) \models S \wedge \text{music} \ \Rightarrow \\ \exists t4 \geq t3 \ \text{state}(\gamma, t4, \text{internal}) \models s2] \\ \forall t4 \ [\ \text{state}(\gamma, t4, \text{internal}) \models s2 \ \Rightarrow \\ \exists t1, t2, t3 \ \ t1 \leq t2 \leq t3 \leq t4 \ \& \\ \text{state}(\gamma, t1, EW) \models \neg S \wedge \neg \text{music} \ \& \\ \text{state}(\gamma, t2, EW) \models \neg S \wedge \text{music} \ \& \\ \text{state}(\gamma, t3, EW) \models S \wedge \text{music}] \end{aligned}$$

This comes close to the transitions indicated by Damasio: *the proto-self exists at the inaugural instant - an object comes into sensory representation - the proto-self has become modified by the object.*

The above relational specification is a first-order representation in the sense that it refers to external states of world and body, whereas Damasio’s second-order

representation refers to internal states (other, first-order, representations) of the proto-self. The relational specification given above only works for body loops, not for ‘as if body loops’. A relational specification that comes more close to Damasio’s formulation, and also works for ‘as if body loops’ is the following (**RSP**):

$$\begin{aligned} \forall t1, t2, t3 \ [\ t1 \leq t2 \leq t3 \ \& \\ \text{state}(\gamma, t1, \text{internal}) \models \neg \text{sr}(S) \wedge \neg \text{sr}(\text{music}) \ \& \\ \text{state}(\gamma, t2, \text{internal}) \models \neg \text{sr}(S) \wedge \text{sr}(\text{music}) \ \& \\ \text{state}(\gamma, t3, \text{internal}) \models \text{sr}(S) \wedge \text{sr}(\text{music}) \ \Rightarrow \\ \exists t4 \geq t3 \ \text{state}(\gamma, t4, \text{internal}) \models s2] \\ \forall t4 \ [\ \text{state}(\gamma, t4, \text{internal}) \models s2 \ \Rightarrow \\ \exists t1, t2, t3 \ \ t1 \leq t2 \leq t3 \leq t4 \ \& \\ \text{state}(\gamma, t1, \text{internal}) \models \neg \text{sr}(S) \wedge \neg \text{sr}(\text{music}) \ \& \\ \text{state}(\gamma, t2, \text{internal}) \models \neg \text{sr}(S) \wedge \text{sr}(\text{music}) \ \& \\ \text{state}(\gamma, t3, \text{internal}) \models \text{sr}(S) \wedge \text{sr}(\text{music}) \] \end{aligned}$$

This is a relational specification in terms of other representations (sr(music), sr(S)), and therefore a second-order representation. It has no direct reference to external states anymore. However, indirectly, via the first-order representations sr(music) and sr(S) it has references to external states.

Verification

In previous sections, local, executable dynamic properties were addressed, and simulation based on these properties was discussed. After that, dynamic properties to describe representational content of internal states are introduced. These dynamic properties are of a *global* nature. Another example of a more global property is the following:

$$\text{OP1} \ \text{music} \leftrightarrow s2$$

Informally, this property states that the presence of music eventually leads to the birth of core consciousness (s2). This can be considered as a global property because it describes dynamic of the overall process, whereas the local properties (LP’s) presented earlier described basic steps of the process. For both types of global properties (i.e., dynamic property OP1 and the properties specifying representational content), an important issue is *verification*. In other words, are these global properties satisfied by the simulation model? Therefore, the global properties have been formalised, and verification has been applied in two ways: by *automated checks* and by establishing *logical relationships*.

For the first type of verification, a software environment has been developed that enables checking dynamic properties specified in TTL against simulation traces. This software environment takes a dynamic property and one or more (empirical or simulated) traces as input, and checks whether the dynamic property holds for the traces. Using this environment, the global properties mentioned above have been automatically checked against traces like depicted in Figure 3. All these checks turned out to be successful, which validates (for the given traces at least) our choice for the representational content of the internal state properties.

A second way of verification is to establish logical relationships between global properties and local properties. This has been performed in a number of cases. For example, to relate OP1 to local properties, intermediate properties

were identified in the form of the following milestone properties that split up the process in three phases:

MP1(MtoE) music \leftrightarrow sr(music) & sr(music) \leftrightarrow S
MP2(EtoF) S \leftrightarrow sr(S)
MP3(FtoFF) **RSP** (see previous section)

For the milestone properties the following relationships hold (for simplicity neglecting ‘as if body loops’):

MP1(MtoE) & MP2(EtoF) & MP3(FtoFF) \Rightarrow OP1
 LP0 & LP1 & LP2 & LP3 \Rightarrow MP1(MtoE)
 LP4 & LP5 \Rightarrow MP2(EtoF)
 LP7 & LP8 & LP9 \Rightarrow MP3(FtoFF)

Such logical relationships between properties can be very useful in the analysis of traces. For example, if a given trace that is unsuccessful does not satisfy milestone property MP2, then by a refutation process it can be concluded that the cause can be found in either LP4 or LP5. In other words, either the sensor mechanism fails (LP4), or the sensory representation mechanism fails (LP5).

Discussion

The chosen modelling approach describes temporal dependencies in processes at a neurological, not symbolic level. To avoid complexity the model is specified at an abstract level. From the available approaches to representational content from Philosophy of Mind, the causal/correlational approach is not applicable, but Kim’s relational specification approach, that allows more complex temporal dependencies, is applicable. Using this approach, claims on representational content made by Damasio have been formalised and supported using two types of verification techniques.

Furthermore, an interesting observation that has been made on the basis of the formalisation was that the model predicted the possibility of ‘false core consciousness’: core consciousness that is attributed to the ‘wrong’ stimulus. To explain this phenomenon, suppose that two stimuli occur, say x1 and x2, where x2 is subliminal and unnoticed. Then, it could be the case that x2 provokes emotional responses, whilst the conscious feeling that arises is attributed to x1 instead of x2. In terms of our model, this can be simulated by first introducing a subliminal stimulus that yields emotion S (e.g., a cold breeze) followed by the stimulus music. In that case, the conscious feeling would incorrectly be attributed to the music. In personal communication with Antonio Damasio, the existence of this predicted false core consciousness was confirmed. Although this is not a proof about the validity of the model, it indicates that this type of modelling can be used to derive interesting predictions.

From a philosophical perspective, the paper contributes a case study for representational content which is more down-to-earth than the science fiction style thought experiments that are common in the literature on Philosophy of Mind (e.g., Kim, 1996). In addition, the type of representation is more sophisticated than the usual ones essentially addressing sensory representations induced by observing (a snapshot of) a horse or a tomato. Interesting further work in

this area is to analyse various arguments given in this literature by applying them to this example.

The analysis approach that is applied in this paper to model Damasio’s theory of consciousness, has previously been applied to complex and dynamic cognitive processes other than consciousness, such as the interaction between agent and environment (Bosse, Jonker & Treur, 2005). In a number of these cases, in addition to simulated traces, also empirical (human) traces have been formally analysed. Using this approach, it is possible to verify global dynamic properties (e.g., specifying the representational content of internal states) in real-world situations.

For recent work in the area of emotion and consciousness, the interested reader is referred to (Prinz and Chalmers, 2004, Chapter 3), which gives an account for emotions as embodied representations of “core relational themes” such as danger and obstruction.

Acknowledgments

The authors are grateful to Antonio Damasio for his valuable comments upon their questions, and to an anonymous referee for some suggestions for improvements of an earlier version of this paper.

References

- Bosse, T., Jonker, C. M., van der Meij, L. & Treur, J. (2005). LEADSTO: a Language and Environment for Analysis of Dynamics by SimulaTiOn. In: Eymann, T. et al. (eds.), *Proceedings of the Third German Conference on Multi-Agent System Technologies, MATES'05*. Lecture Notes in AI, vol. 3550. Springer Verlag, pp. 165-178.
- Bosse, T., Jonker, C.M. & Treur, J. (2005). *Representational Content and the Reciprocal Interplay of Agent and Environment* In: Leite, J., Omicini, A., Torroni, P., and Yolum, P. (eds.), *Proc. of the Second International Workshop on Declarative Agent Languages and Technologies, DALT'04*, Springer Verlag, pp. 61-76.
- Damasio, A. (2000). *The Feeling of What Happens: Body, Emotion and the Making of Consciousness*. MIT Press.
- Damasio, A., Grabowski, T.J., Bechara, A., Damasio, H., Ponto, L.L.B., Parvizi, J. & Hichwa, R.D. (2000). Subcortical and cortical brain activity during the feeling of self-generated emotions. *Nature Neuroscience*, vol. 3, pp. 1049-1056.
- Jonker, C.M. & Treur, J. (2002). Compositional Verification of Multi-Agent Systems: a Formal Analysis of Proactiveness and Reactiveness. *International Journal of Cooperative Information Systems*, vol. 11, pp. 51-92.
- Jonker, C.M., Treur, J. & Wijngaards, W.C.A. (2003). A Temporal Modelling Environment for Internally Grounded Beliefs, Desires and Intentions. *Cognitive Systems Research Journal*, vol. 4, pp. 191-210.
- Kim, J. (1996). *Philosophy of Mind*. Westview Press.
- Parvizi, J. & Damasio, A. (2001). Consciousness and the brain stem. *Cognition*, vol. 79, pp. 135-159.
- Prinz, J.J. & Chalmers, D.J. (2004). *Gut Reactions: A Perceptual Theory of Emotion* (Philosophy of Mind (Hardcover)), Oxford University Press.